PROJECT_BIG DATA _CLASSIFICATION PREDICTION MODELS_MUWANI_ROBSON

BACKGROUND SUMMARY

This project report summarizes supervised machine learning task based on SPARK APACHE(unified analytics for big data processing). Smoking-related illnesses cost the United States and Canada hundreds of billions of dollars a year in health care expenditures and lost productivity, and claim hundreds of thousands of lives. Given the enormous medical and economic toll of smoking, it is not surprising that insurance companies favor charging smokers higher rates for health insurance to provide them with an incentive to stop smoking. The task is based on a classification problem to predict smokers or nonsmokers from healthcare insurance database. As such, the aim is to use Insurance data set to predict smokers given a set of input parameters. The data set was an open source data from ML learning repository. The set predictors were based on **: age, gender, bmi, family size and charges.**

Classification Dataset				
Sr No	**Dataset Name**	**Description**	**Rows**	**Columns**
1	**Insurance data set**	Based on Predictors (age, sex, bmi, family size, region and charges we are predicting whether a person is a smoker or nonsmoker using insurance database information.	1338	6
2	**Attribute Information**	-age[numeric data type] -sex[male, female, string encoded to numeric] -bmi[numeric data type] -family size[integer data type] -region[string, encoded to numeric] -charges[float/double data type] -smoker[string encoded to 0(None smoker and Yes(1) for smokers	1338	6

MODEL RESULTS, ACCURACY & CROSS VALIDATION

CLASSIFICATION			
		Logistic Regression(LR)	**Random Forests(RF)**
Insurance Data	**Accuracy**	**0.92096**	**0.96669**
	Recall	**0.92096**	**0.96536**
	Precision	**0.93989**	**0.96669**

CONCLUSIONS & INFERENCES

- ✓ Random forest proved to be superior model in predicting the target class variable(smokers or nonsmokers).
- ✓ Model accuracy for logistic regression was 92% and acceptable in predicting the class variable also.
- ✓ Random forests outperformed logistic regression on both precision, model score above 96%, thus the measure proportion of positive identifications(smokers or nonsmokers) was correct, smokers correctly predicted as smokers while in logistic the correct prediction was at 92%.
- ✓ In terms of recall as a measure of sensitivity, i.e. What proportion of actual positives(smokers/non smokers) was identified correctly ,Random Forest outperformed Logistic Regression. The model score above 96% times better than 94% in logistic regression model.
- ✓ Logistic regression posted an improvement in precision(94%) score after applying Gridsearch on the two models by tuning for the best hyperparameters.

MODEL RECOMMENDATIONS & SUGGESTIONS FOR FURTHER IMPROVEMENT

- ✓ Recommend use of Random forests as the best model in predicting the label class(smokers or non smokers), although logistic regression was also not a bad model.
- ✓ Need to watch for overfitting in the two models by ensuring the proper models tuning in selecting all the best hyperparameters through GridsearchCV and regularization of the two models as we apply on different data sets.
- ✓ Recommend use of balanced data for the target class: smokers and non-smokers for best model results. This can be achieved through up sampling or down sampling and k fold cross validation.
- ✓ **Improve model by adding more data**-for models with high variance high variance of predictions it is ideal to increase the training set size. Try increasing your sample by providing new data, which could translate into new cases or new features
- ✓ Algorithm tuning and use of multiple algorithms improves outcome of model learning process, Transforming before multilevel modelling (thus attempting to make coefficients more comparable, thus allowing more effective second-level regressions, which in turn improve partial pooling
- ✓ **Use of ensemble-**apply bagging (to reduce variance) and boosting (adjust the weights based on the last classification)
- ✓ **Feature engineering**-apart from transformations, creating new variables out of existing variables is also very helpful to create new features that improve the model's performance. Every new feature can make guessing the target response easier like in Principle Component Analysis (PCA). Automatic feature creation is possible using polynomial expansion or the support vector machines class of machine learning algorithms. Support vector machines can automatically look for better features in higher-dimensional feature spaces in a way that is both computationally fast and memory optimal.

……………………………………………………………………………***END***--